Page Boundary Extraction of Bound Historical Herbaria

Krishna Kumar Thirukokaranam Chandrasekar[a] and Steven Verstockt[b]
IDLab, IMEC-Ghent University, Ghent, Belgium
{*krishnakumar.tc, steven.verstockt*}*@ugent.be*

Keywords: Page Boundary Detection, Hinge Detection, Border Noise, Digitization, Historical Image Processing, Herbaria Books.

Abstract: When digitizing bound historical collections such as herbaria it is important to extract the main page region so that it could be used for automated processing. The thickness of the herbaria books also gives rise to deformations during imaging which reduces the efficiency of automatic detection tasks. In this work we address these problems by proposing an automatic page detection algorithm that estimates all the boundaries of the page and performs morphological corrections in order to reduce deformations. The algorithm extracts features from Hue, Saturation and Value transformations of an RGB image to detect the main page polygon. The algorithm was evaluated on multiple textual and herbaria type historical collections and obtains over 94% mean intersection over union on all these datasets. Additionally, the algorithm was also subjected to an ablation test to demonstrate the importance of morphological corrections.

1 INTRODUCTION

Since the early 1990s, libraries and museums have conducted multiple digitization initiatives with cultural heritage documents and scientific resources on regular basis to ensure restoration and lasting preservation of historical collections. This is to protect them from further degradation caused by repetitive handling. Exponential growth in high quality image capturing devices induced by the enormous amount of rich historical collections (that are yet to be uncovered) has further led to a raising interest in historical document image analysis in recent times. Indeed, an important need has also emerged to develop automated tools to process and enrich these collections to facilitate better access to the preserved archives.

In addition to textual documents and records such as books, student registers or death records that are normally digitized on a large scale, there are various types of bound historical herbaria that preserve the rich horticulture of a region, that should also be digitized. In these herbaria, plants are collected, dried and stored in often difficult circumstances so that it could be used as a reference material decades later. It is important to preserve these records with technical proficiency but at the same time make it available to readers easily. Though modern scanners offers solutions for preserving these information, scanned materials are not always correctly oriented along the coordinate axes. In certain instances, the visual information is corrupted by external border noise. Sometimes it also introduces significant variations in the page by inducing noise such as unimportant objects, slightly rotated, imaged under different viewing angles and perspective or blended with a scene background. Such noises interrupt the document analysis algorithms, that in turn affect the efficiency of the overall scanning procedure. While removing background is feasible using simple segmentation techniques, other types of border noises are more challenging.

The digitization of books and bound herbaria also suffer from deformation and warping due to the thickness of the collections. This poses a huge problem for automated detection tasks. For textual books, such warping effect decreases the text recognition process to a great extent. A similar problem persists for herbaria, wherein warping and deformation affects the original shape and texture of the leaf specimen. This directly affects historical plant phenotypeing and experiments involving learning the evolution of shape of leaves for which these digitized herbaria could be greatly useful. Therefore it makes it necessary to accurately detect the boundaries of the page, including the edge that is usually shared by the neighbouring page. This center portion of the page that is usually caused due to binding is called *hinge* in book anatomy

[a] https://orcid.org/0000-0003-4385-915X
[b] https://orcid.org/0000-0003-1094-2184

Figure 1: A sample scan image of a historical herbaria during the digitization process.

based literature. The following terminology would be used in the rest of the paper to refer that edge. The detection of hinge gets increasingly tougher for hand binded books since the edge would not be straight any more.

A number of approaches have been developed to remove border noise (e.g., [(Bukhari et al., 2012), (Chakraborty and Blumenstein, 2016b), (Fan et al., 2002), (Shafait and Breuel, 2010a), (Shafait and Breuel, 2010b)]). However, as noted in (Chakraborty and Blumenstein, 2016a), most of the prior methodologies make fixed assumptions that holds good only for textual pages. Some of these assumptions are consistent text size, absolute location of border noise, straight text lines, and distances between page text and border. These assumption don't hold good for herbaria type images (e.g. Figure 1) where such a consistent pattern is not followed. Therefore it is necessary to have an algorithm that makes use of only page and book based features such as colour, intensity, illumination and texture to detect the page boundaries.

In order to address the above concerns we propose a multi step page detection algorithm that detects variations in brightness and the distribution of the colour in an image to estimate the page boundary. The algorithm is based on the HSV colour model since it provides additional intensity and colour depth features that can be utilized to better localize the page boundaries. A novel hinge detection algorithm is also proposed that can be used to specifically localize and extract the center portion(hinges) of the book.

The remaining paper is organized as follows. Section 2 explains the algorithms in detail and reasons the approaches with examples. The data sets used are elaborated in Section 3. Section 4 discusses the initial results, while Section 5 concludes the paper by proposing possible future work directions.

2 METHODOLOGY

Page detection is considered as the process of finding pixels and regions in an image that constitutes a page. Within the domain of historical digitization, page detection is predominantly applied for pre processing of documents before hand written text detection and recognition tasks, line and character detection and segmentation of historical pictures.

As shown in Figure 2, the proposed pipeline for page detection begins with the pre processing of images. The images are rotated and aligned such that the longest edge is maintained as its height. The core methodology of the proposed algorithm can be sub divided into two main steps namely book extraction and hinge region detection. The book extraction step filters background noise and extracts the main book region while the hinge region detection step detects the hinges and extracts the main page region. Finally, morphological transformations are performed on the extracted page in order to reduce deformations.

Since our algorithm is directed primarily towards historical documents, we would be dealing with books and herbaria from the 19th and the early 20th century. Due to the factor of aging, a large majority of these books possess a distinctive texture and colour which could be used for detection. Additionally, it is also necessary to represent the image using a colour space (Sanchez-Cuevas et al., 2013; Albiol et al., 2001) in which some of the colour channels are invariant or at least insensitive to lighting changes, such as the H and S channel in HSV (James, 2013). In HSV colour space, the Hue (H) and Saturation (S) values represent the colour information of the image. The Value (V) on the other hand is the measure of light intensity and denotes the extent of the colour's brightness of the image. A range of Hue, Saturation and Value values of HSV can be used as important features for page boundary detection as further explained below.

Figure 2: The proposed page detection pipeline.

2.1 Book Extraction

The Saturation and Value components of HSV colour space have been adopted as the major features for background filtering and initial boundary prediction. Figure 3 provides a visual comparison of the original image with their Saturation and Value transformed images. The transformed images can directly be used for filtering and segmentation of the page to remove external and surrounding background noises.

Figure 3: The first image from the left is the original image and the following two images are its corresponding Saturation and Value transformed images. It can be seen that the hinge is more prominently perceivable from the saturation image (middle image) which we would use to localize the hinge edge automatically.

Based on the average histogram of the Hue values for a random selection of 100 historic images from 6 different collections, it was found that more than 85% of the pixel's expected dominant colours lie between 5° and 65° of the Hue colour wheel. This seems to be logically acceptable since the dominant colours between the specified range are red and yellow and the majority of historic images has a high probability of having a slight yellow or brown tint due to the ageing factor of the document. The average histogram of Values shows a much more wider distribution. Yet, on further introspection of the average histograms, it has been witnessed that all the historic books taken into consideration, falls between the range of 30 - 93%. Based on the observed values for Hue and Value, book mask B_{mask} can be estimated for the image based on Equation 1.

$$B_{mask} = \left\{ \begin{array}{ll} 1, & \begin{array}{c} 5^\circ < H < 65^\circ \\ 30\% < V < 93\% \end{array} \\ 0, & otherwise \end{array} \right\} \quad (1)$$

An example of the filtered mask obtained using HSV filtering is shown in Figure 4. Based on calculating the largest contour from the mask, the background and border noise can be eliminated and the book can be extracted by calculating the convex hull points of the contour (Goodrich et al., 2009).

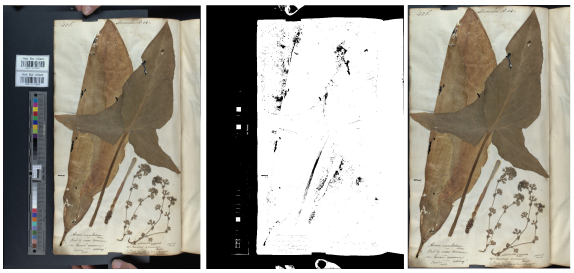


Figure 4: An example of book extraction with the corresponding segmentation mask generated based on HSV filtering.

2.2 Hinge Region Detection

The HSV based filtering and segmentation results in the elimination of background noise, yet it is not sufficient enough to predict the hinge of the page. This is because the hinge is usually shared by the neighbouring page and therefore just colour based techniques are inadequate.

Based on human perception, even though the fourth boundary of a page in a book is shared by the neighbouring pages, it is still possible to locate this boundary because the area around the boundary tends to be darker. Logically this is because, the amount of light that could reach the hinge is lower due to binding. The Saturation value (S) signifies the amount of white light that needs to be mixed with the Hue and therefore could be used as a feature to detect hinges. In the center image of Figure 3 it is seen that the region around the hinge is brighter than the other parts of the page. Therefore, the Saturation values could be used for detecting the hinge.

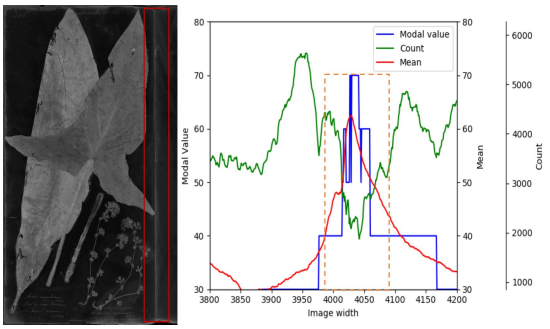


Figure 5: The plot shows the mean, mode and the mode count values for the columns of the image enclosed within the red bounding box. The orange dashed box shows the region where the actual boundary lies.

The plot in Figure 5 depicts the correlation between the mean, mode and modal count values for a selected region around the boundary of the image. Since our focus here is predicting a longitudinal boundary, the mean and modal values are calculated for every column of the image Saturation values. There is an evident Gaussian behaviour for the mean and modal values around the boundary region, which can be further justified by the plot in Figure 6.

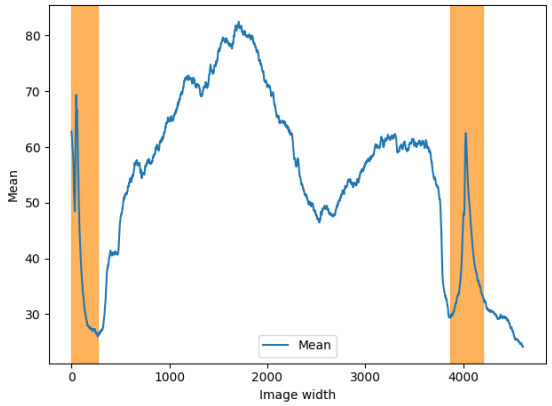


Figure 6: The plot shows the mean value of the columns for the image in Figure 4. The orange strips indicate the region of the actual boundary.

The above mentioned phenomenon is made use of to localize and detect the hinges with near pixel level accuracy. To begin with, the peaks and troughs are estimated using a function that can find all local maxima by simple comparison of neighbouring values (Virtanen et al., 2019). The scipy implementation *find_peaks* has been utilized to detect the peaks and troughs. The parameter prominence had a greater effect on detecting the peaks and in our experiments prominence=10 was chosen as it performed the best for majority of images. Finally hinge regions are estimated based on the following conditions:

- Based on the peaks and trough values of the mean, the peaks with the steepest slopes are selected.
- For the selected slopes, the mode and mode count values are cross verified. Based on the plot in Figure 5 it can be inferred that the modal count and mean values are inversely correlated. Thus only those peaks that satisfy this condition are selected.
- Since the book is initially aligned in such a way that the hinge is always to the right side of the image, peaks present in the first half of the image are eliminated. It is also possible to automatically select the peaks based on the location of the contour but since the image is rotated and aligned during the pre processing, we follow this methodology.
- From the remaining list of peaks, the peak with the highest prominence is chosen. The hinge region is derived as the region between the two local minima before and after the peak as indicated by

the orange strip in Figure 6.

Note: Since the dimensions of the page do not change over a book, it is also possible to obtain them before the start of the digitization process. In case if the dimensions of the page are already known, the peaks can be selected based on the width/height ratio of the book and detect the page in the image.

2.3 Page Detection

Based on the obtained book mask and hinge region, the overall page is detected as follows:

- For the selected hinge region, the Saturation values (S) are applied a threshold and transformed into a mask based on the following criteria:

  Let h_r be the selected hinge region. Then

  $$hs_{mean} = mean(S[h_r])$$
  $$hs_{max} = max(S[h_r])$$

  $$h_{mask} = \left\{ \begin{array}{ll} 0, & hs_{mean} < hs_r < hs_{max} \\ 1, & otherwise \end{array} \right\} \quad (2)$$

- The calculated h_{mask} is combined with the original book mask as follows:

  $$B_{mask}[h_r] = h_{mask}$$

- The page region P is estimated by applying hull points to the largest contour (Goodrich et al., 2009). The largest contour is calculated as follows.

  $$P = max_{area}[B_{mask}]$$

The overall algorithm for page detection is elaborated in Algorithm 1.

2.4 Morphological Correction

The estimated page region is expected to be a rectangle or square for a normal book based on its dimensions. Yet, the obtained page suffers a mismatch due to a number of external factors. In order to compensate for this mismatch, the detected page has to be interpolated and transformed such that their shapes match. The morphological correction is performed as follows:

- For the estimated page region P, enclosing bounding box is calculated. The enclosing bounding box would be the reference bounding box R. In case if the book dimensions are known, the dimensions are used to estimate the reference bounding box. The morphological correction is performed if the area of P is less than the area of R.
- In order to learn the shape representations, i equiv distant points are chosen along every side s of P and reference bounding box R. The value is dependent on the actual size and resolution of the image. In our algorithm, 12 points were chosen on each side. 12 was chosen since for i =12, the selected points were neither too close nor too far away resulting in better interpolation results.
- P is interpolated between P_{si} and R_{si} and the resulting co-efficients are re mapped on to R_{si}. In our algorithm, cubic interpolation was used since it performed better than linear interpolation.

The implementation of morphological correction is demonstrated in Figure 9.

Algorithm 1: Page Detection.

Require: An image of a book page
  Convert RGB to HSV
  Generate book masks using HSV filtering.
  if book is not portrait **then**
    Rotate the mask by 90° and straighten the image
  else
    Straighten the image
  end if
  $fullpage \leftarrow$ rectangle enclosing the largest contour
  $Mean \leftarrow$ mean of full page[axis=0]
  $Mode, Count \leftarrow$ Mode of full page[axis=0]
  $Peaks, troughs \leftarrow$ Find peaks and troughs for mean and count
  Detect Hinge region
  if Hinge region **then**
    Threshold the region
    Attach the region to the Book mask
  end if
  Extract Page P based on the largest contour of the mask
  Perform morphological correction

3 DATASET

The main datasets that were used for our experiments were herbaria books of three different and prominent Belgian botanist from the late 18th and early 19th century, Charles Van Hoorebeke, Aimé Mac Leod and Julius Mac Leod . There are 78 books of Charles Van Hoorebeke with 20-40 single sided specimen per

Table 1: Evaluation of the proposed algorithm on multiple datasets.

Dataset	Precision	Recall	IoU
Mac Leod	0.963	0.951	0.966
Charles Van Hoorebeke	0.935	0.948	0.94
cBAD	0.937	0.973	0.942

(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 7: Sample results of the page detection algorithm. (a) is a page from Charles Van Hoorebeke, (b) and (d) are from Mac leod herbaria. The rest are from cBAD baseline - dataset with different complexity obtained from 7 different archives.

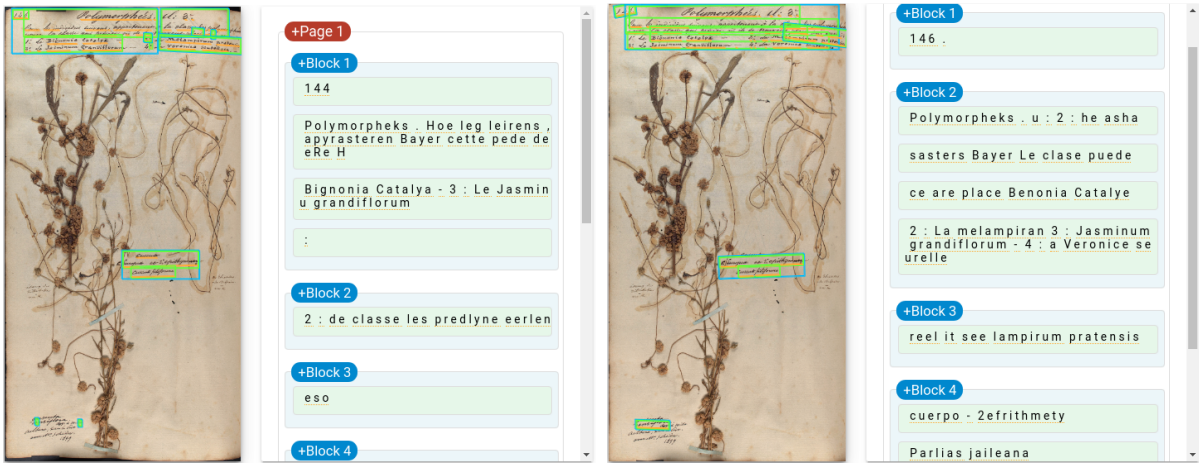
book. All these books possess a similar structure (e.g. Figure 7(a)). The books of Julius and Aime Mac Leod are a bit more complex with approximately 200 pages each and with multiple specimens per page (e.g. Figure 7(b), 7(d)). The page detection algorithms were initially developed and parameters were chosen based on randomly selected pages from these books. The final algorithm was evaluated on the rest of the pages.

In order to evaluate the generalization of the algorithm, the same algorithm was also evaluated on the ICDAR 2019 Competition Baseline Detection (cBAD) dataset (Markus et al., 2019). The dataset consist of documents with varying levels of layout complexity extracted from 7 archives (bottom row in Figure 7). There were documents from two tracks and we used a random subset of documents from both of the tracks.

3.1 Evaluation

Precision, recall and IoU (Intersection over Union) metrics are normally used to evaluate the efficiency of the boundary detection algorithms. These metrics should be used in order to compare performance of the algorithm with similar models. As proposed in (Tensmeyer et al., 2017) we also evaluate the page detection algorithm using manually labelled polygons.

Normally a page edge could be a few pixels thick and it is important to evaluate how much of the edge actually falls within this region. A pixel level comparison is therefore performed for the detected boundary to evaluate the closeness of the estimation with the



(a) (b)

Figure 8: Comparison of text detection and recognition results using Google Cloud Vision API. (a) depicts the recognition results for original non-corrected page while (b) depicts the results for a morphologically corrected page. It can be seen that the detection and grouping of text (light blue bounding box), number recognition (the right page number is *146*) and text recognition (*la melampirum, jasminum grandiflorum, veronice seurellera*) results are much improved in (b).

ground truth. For this type of evaluation, the page boundary regions were manually annotated for 50 images of variable complexity as shown in Figure 7 bottom row. Finally, the test examples were also manually evaluated using two human participants.

4 RESULTS AND DISCUSSION

Table 1 shows the precision, recall and IoU scores for the different datasets. It could be seen that the overall IoU score is minimum 94% over all the datasets combined. Even though the following results were comparable with the models of (Tensmeyer et al., 2017), as shown in Figure 7, the page boundary mask could be used for further processing of the page.

Since a precise shape of the page boundary polygon could be obtained, this feature can be used for reducing the deformation of the page. Figure 9 showcases one such use case for flattening of the page to reduce deformation using simple morphological transformation. For this task, the numpy implementation *griddata* was used to learn the current representation of the page and was transformed into the original representation (before deformation)using the opencv geometric image transformation function *remap*. Normally, text lines would be detected to reduce deformation. But, for herbaria type collections, where text patterns would be limited, it would be hard to use the text line features. In those scenarios, the page boundaries can be used as features to reduce deformation.

(a) (b)

Figure 9: An example of morphological correction for flattening the page in order to reduce deformation. (a) is the original image with chosen hull points for flattening and (b) is the subsequent flattened image.

4.1 Ablation Study

An ablation study was performed to evaluate the importance of morphological correction. To do so, the page samples before and after morphological transformations of the Mac Load test data set were chosen and text detection / recognition was applied. The Google Cloud Vision API was used to obtain text detection and recognition. The results of text detection and recognition were quantitatively observed. The

number of right predictions for both text detection and recognition tasks were manually verified and counted for both the samples of the image. Since the image quality between the two images is similar and the same text recognition model setting is used, it makes the obtained results comparable. It was observed that the morphological correction improved both the localization and prediction of hand written text by 25% on average for each image. It was also noticed that page numbers, headings and text that were close to the boundaries had much better results than before. An example of the result is shown in Figure 8.

5 CONCLUSIONS

A novel page detection algorithm has been presented which eliminates border noise by segmenting the main page region from the rest of the image. The importance of using HSV colour model for historical document processing was elaborated. With less assumptions, it was showed that the page detection could also work for complex page structures. It was also demonstrated that the detected page polygon could be used as a feature for reducing deformation. Finally, the page with reduced deformations was proved to perform better in automatic text detection tasks.

ACKNOWLEDGEMENTS

The research activities described in this paper were funded by The Department of Culture, Youth & Media, Flanders (Belgium) for the *Flore de Gand* project.

REFERENCES

Albiol, A., Torres, L., and Delp, E. J. (2001). Optimum color spaces for skin detection. *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)*, 1:122–124 vol.1.

Bukhari, S. S., Shafait, F., and Breuel, T. M. (2012). Border noise removal of camera-captured document images using page frame detection. In Iwamura, M. and Shafait, F., editors, *Camera-Based Document Analysis and Recognition*, pages 126–137, Berlin, Heidelberg. Springer Berlin Heidelberg.

Chakraborty, A. and Blumenstein, M. (2016a). Marginal noise reduction in historical handwritten documents – a survey. *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 323–328.

Chakraborty, A. and Blumenstein, M. (2016b). Preserving text content from historical handwritten documents. *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 329–334.

Fan, K.-C., Wang, Y.-K., and Lay, T.-R. (2002). Marginal noise removal of document images. *Pattern Recognition*, 35(11):2593 – 2611.

Goodrich, B., Albrecht, D., and Tischer, P. (2009). Algorithms for the computation of reduced convex hulls. In Nicholson, A. and Li, X., editors, *AI 2009: Advances in Artificial Intelligence*, pages 230–239, Berlin, Heidelberg. Springer Berlin Heidelberg.

James, S. P. (2013). Face image retrieval with hsv color space using clustering techniques.

Markus, D., Florian, K., and Basilis, G. (2019). ICDAR 2019 Competition on Baseline Detection (cBAD).

Sanchez-Cuevas, M. C., Aguilar-Ponce, R. M., and Tecpanecatl-Xihuitl, J. L. (2013). A comparison of color models for color face segmentation. *Procedia Technology*, 7:134 – 141. 3rd Iberoamerican Conference on Electronics Engineering and Computer Science, CIIECC 2013.

Shafait, F. and Breuel, T. (2010a). A simple and effective approach for border noise removal from document images. pages 1 – 5.

Shafait, F. and Breuel, T. (2010b). A simple and effective approach for border noise removal from document images. pages 1 – 5.

Tensmeyer, C., Davis, B., Wigington, C., Lee, I., and Barrett, B. (2017). Pagenet: Page boundary extraction in historical handwritten documents. pages 59–64.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E. W., Vand erPlas, J., Laxalde, D., Perktold, J., Cimrman, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and Contributors, S. . . (2019). SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. *arXiv e-prints*, page arXiv:1907.10121.